Bayesian modelling of Dupuytren disease by using Gaussian copula graphical models

Mohammadi, Abdolreza; Abegaz, Fentaw; van den Heuvel, Edwin; Wit, Ernst C.

Published in:
Journal of the Royal Statistical Society. Series C: Applied Statistics

DOI:
10.1111/rssc.12171

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2017)
66, *Part* 3, *pp.* 629–645

Bayesian modelling of Dupuytren disease by using Gaussian copula graphical models

Abdolreza Mohammadi,

Tilburg University, The Netherlands

Fentaw Abegaz,

University of Liege, Belgium

Edwin van den Heuvel

Eindhoven University of Technology, The Netherlands

and Ernst C. Wit

University of Groningen, The Netherlands

[Received October 2015. Final revision June 2016]

Summary. Dupuytren disease is a fibroproliferative disorder with unknown aetiology that often progresses and eventually can cause permanent contractures of the fingers affected. We provide a computationally efficient Bayesian framework to discover potential risk factors and investigate which fingers are jointly affected. Our Bayesian approach is based on Gaussian copula graphical models, which provide a way to discover the underlying conditional independence structure of variables in multivariate data of mixed types. In particular, we combine the semiparametric Gaussian copula with extended rank likelihood to analyse multivariate data of mixed types with arbitrary marginal distributions. For structural learning, we construct a computationally efficient search algorithm by using a transdimensional Markov chain Monte Carlo algorithm based on a birth–death process. In addition, to make our statistical method easily accessible to other researchers, we have implemented our method in C++ and provide an interface with R software as an R package BDgraph, which is freely available from http://CRAN.R-project.org/package=BDgraph.

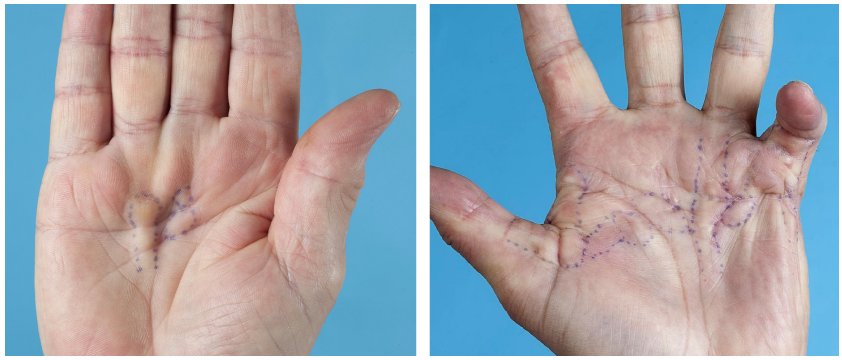
Keywords: Bayesian inference; Bayesian model averaging; Birth–death process; Dupuytren disease; Gaussian copula graphical models; Risk factors

1. Introduction

Dupuytren disease is a hereditary disorder that affects people world wide. It is, however, more prevalent in people with northern European ancestry (Bayat and McGrouther, 2006). The disease is an incurable fibroproliferative disorder that alters the palmar fascia of the hand and may cause progressive and permanent flection contracture of the fingers. Initially, skin pittings and subcutaneous nodules appear in the palm; Fig. 1(a). At a later stage, cords appear that connect the nodules and may contract the fingers into a flexed position; Fig. 1(b). Contractures can arise in a single ray or in multiple rays. The disease mostly appears on the ulnar side of the hand, i.e.

Address for correspondence: Abdolreza Mohammadi, Department of Methodology and Statistics, Tilburg University, Tilburg, The Netherlands.
E-mail: a.mohammadi@uvt.nl

 0035–9254/17/66629
Published by John Wiley & Sons Ltd on behalf of the Royal Statistical Society.



(a) (b)

Fig. 1. (a) A patient with a mild form of Dupuytren disease whose fingers have not been affected by the disease (the palmar nodules and small cords have no signs of contracture) and (b) a patient with Dupuytren disease whose fingers have been affected by the disease (image provided by D. C. Broekstra and P. M. N. Werker at the University Medical Centre Groningen)

it affects the pinky and ring fingers most frequently (see Fig. 3 in Section 3). The only available treatment is surgical intervention. Although much is known about the disease, the questions arising are as follows.

(a) What variables affect the disease and in what way?
(b) Should surgical intervention focus on single or on multiple fingers?

The first is an epidemiological question; the second is a clinical question.

Empirical research has described the patterns of occurrence of Dupuytren disease in multiple fingers, Meyerding (1936) stated that the disease mainly affects the pinky and ring fingers; at the later stage also the middle and index fingers and the thumb may be involved. Tubiana *et al.* (1982) found that the disease barely affects the radial side alone, and that the radial effect is often associated with an affected ulnar side. More recently, Lanting, Nooraee, Werker and van den Heuvel (2014), using a multivariate ordinal logit model, suggested that the middle finger is substantially correlated with other fingers on the ulnar side, and disease in the thumb and index finger are correlated. They took into account age and sex, and tested for hypotheses of independence between groups of fingers. However, so far, no serious multivariate analysis of the disease has been performed taking into account potential risk factors.

Essential risk factors of Dupuytren disease include both phenotypic and genotypic factors (Shih and Bayat, 2010), such as genetic predisposition and ethnicity, as well as sex and age. However, it is unclear whether Dupuytren disease is a complex oligogenic or a simple monogenic Mendelian disorder. Several lifestyle risk factors (some considered controversial), including smoking, excessive alcohol consumption, manual work and hand trauma, have been linked to the disease (Geoghegan *et al.*, 2004; Godtfredsen *et al.*, 2004). In addition, several diseases, such as *diabetes mellitus* and epilepsy, are thought to affect the severity of Dupuytren disease. However, the role of these lifestyle factors and diseases has not been fully elucidated, and the results of different studies are occasionally conflicting (Lanting, Broekstra, Werker and van den Heuvel, 2014).

In this paper we analyse data collected by the Department of Plastic Surgery of the University Medical Centre Groningen in the north of the Netherlands involving patients who have Dupuytren disease. Both hands of the patients are examined for signs of Dupuytren disease. These are tethering of the skin, nodules, cords and finger contractures in patients with cords. The

severity of the disease is measured by the angles on each of the 10 fingers. Recorded potential risk factors include smoking habits, alcohol consumption, whether participants had performed manual labour during a significant part of their life and whether they had sustained hand injury in the past, including surgery. In addition, information about the presence of Ledderhose diabetes, epilepsy, peyronie, knuckle pad, liver disease and familial occurrence of Dupuytren disease, defined as a first-degree relative with Dupuytren disease, was collected.

The primary aim of this paper is to model the relationships between the risk factors and disease indicators for Dupuytren disease from the viewpoint of multivariate data analysis. In this regard, graphical models (Lauritzen, 1996) provide a potential way to decode the underlying relationships between variables in multivariate data. However, most research effort in the literature has been focused on multivariate normal models; see Mohammadi and Wit (2015), Dobra *et al*. (2011), Yuan and Lin (2007) and Meinshausen and Bühlmann (2006) and their references. A frequentist method for inference of graphical models with mixed variables was introduced in Abegaz and Wit (2015) and Liu *et al*. (2012). We develop a computationally efficient Bayesian statistical method based on Gaussian copula graphical models (GCGMs) for discovering the joint conditional independence structure of binary, ordinal or continuous variables simultaneously.

Our Bayesian framework is based on the GCGMs that were proposed by Dobra and Lenkoski (2011). In GCGMs, the graph selection procedure is embedded inside a semiparametric framework, using the extended rank likelihood (Hoff, 2007). In this paper we design our Bayesian framework for GCGMs on the basis of a computationally efficient search algorithm, using a transdimensional Markov chain Monte Carlo (MCMC) approach based on a continuous time birth–death process (Mohammadi and Wit, 2015, 2016a). The algorithm that Mohammadi and Wit (2015) proposed is concerned with Gaussian graphical models only. The copula approach allows more general data structures of mixed type. Furthermore, our approach can handle missing data without any additional computational effort, if the missingness is completely at random.

In Section 2 we illustrate our Bayesian framework based on GCGMs. In addition, we describe the performance of our method and we compare it with state of the art alternatives. In Section 3 we analyse the Dupuytren disease data set based on our Bayesian method. In this section, first we study potential phenotype risk factors for Dupuytren disease. Second, we analyse the relationship between the severity of Dupuytren disease in pairs of fingers on both hands. The result may help surgeons to decide whether they should operate on one finger or whether they should operate on multiple fingers simultaneously. Finally, we discuss the connections between existing methods and possible future directions.

The program that was used to analyse the data can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. Methodology

Graphical models provide an effective way to describe statistical patterns in multivariate data (Whittaker, 2009; Lauritzen, 1996). In this context undirected Gaussian graphical models are commonly used, since inference in such models is tractable. In such graphical models, the graph structure is characterized by its precision matrix, i.e. the inverse of the covariance matrix: the non-zero entries in the precision matrix show the edges in the conditional independence graph. In the real world, however, data are often non-Gaussian, like the data set that is considered in this paper. For non-Gaussian continuous data, variables can be transformed one

to one to Gaussian latent variables. For discrete variables, however, there is no one-to-one transformation into latent Gaussian distributions. A common approach is to apply an MCMC method to simulate both the latent Gaussian variables and the posterior distributions (Hoff, 2007). Another Bayesian approach is the GCGMs that were developed by Dobra and Lenkoski (2011), in which the sampling algorithm is based on a reversible jump MCMC algorithm and a Cholesky decomposition of the precision matrix. Alternatively, our method implements a birth–death MCMC approach (Mohammadi and Wit, 2015), which has several computational advantages compared with the reversible jump MCMC approach as we show in our simulation examples.

2.1. Gaussian copula graphical models

In graphical models, conditional dependence relationships between random variables are presented as a graph G. A graph $G=(V,E)$ specifies a set of vertices $V=\{1,2,...,p\}$, where each vertex corresponds to a random variable, and a set of edges E. The absence of an edge between two vertices specifies the pairwise conditional independence of these two variables given the remaining variables, whereas an edge between two variables determines the conditional dependence of the variables. In our application, for example, disease risk factors (such as disease factors, alcohol and hand injury) will be the nodes, and dependences between them will be the edges.

Copulas provide a flexible tool for understanding dependence between random variables, in particular for non-Gaussian multivariate data. In our case, we consider 23 variables of mixed type: discrete, binary and ordered categorical variables; see Section 3.1.

Let Y be a collection of continuous, binary, ordinal or count variables with marginal distribution F_j of Y_j and F_j^{-1} its pseudoinverse. For constructing a joint distribution of Y, we introduce a multivariate normal latent variable as follows:

$$Z \sim \mathcal{N}_p(0,\Sigma),$$

where Σ is the correlation matrix. We define the observed data as

$$Y_j = F_j^{-1}\{\Phi(Z_j)\}.$$

A Gaussian copula-based joint cumulative distribution of Y is given by

$$P(Y_1 \leqslant y_1,\ldots,Y_p \leqslant y_p) = \Phi_p[\Phi^{-1}\{F_1(y_1)\},\ldots,\Phi^{-1}\{F(y_p)\}|\Gamma].$$

Our aim is to infer the underlying graph structure of the mixed variables $\mathbf{Y}=(Y^{(1)},\ldots,Y^{(n)})$ implied by n independent draws of latent Gaussian variables $\mathbf{Z}=(Z^{(1)},\ldots,Z^{(n)})$. Following Hoff (2007) and Dobra and Lenkoski (2011), we impose the latent samples $\mathbf{z}$, given the observations $\mathbf{y}$, to belong to the set

$$\mathcal{A}(\mathbf{y}) = \{\mathbf{z}\in\mathbb{R}^{n\times p} : \max\{z_j^{(s)} : y_j^{(s)} < y_j^{(r)}\} < z_j^{(r)} < \min\{z_j^{(s)} : y_j^{(r)} < y_j^{(s)}\}\}. \tag{1}$$

It follows that inference on the latent space can be performed by substituting the observed data $\mathbf{y}$ with the event $\mathbf{z}\in\mathcal{A}(\mathbf{y})$. For a given graph G and precision matrix $K=\Sigma^{-1}$, the extended rank likelihood is defined as

$$P\{\mathbf{z}\in\mathcal{A}(\mathbf{y})|K,G\} = \int_{\mathcal{A}(\mathbf{y})} P(\mathbf{z}|K,G)\,\mathrm{d}\mathbf{z}. \tag{2}$$

In the next sections we develop a Bayesian approach based on the extended rank likelihood given in equation (2).

2.2. Bayesian Gaussian copula graphical models

In the Bayesian framework, we consider the joint posterior distribution

$$P\{K, G|\mathbf{z} \in \mathcal{A}(\mathbf{y})\} \propto P\{\mathbf{z} \in \mathcal{A}(\mathbf{y})|K, G\}\, P(K|G)\, P(G), \tag{3}$$

where $P\{\mathbf{z} \in \mathcal{A}(\mathbf{y})|K, G\}$ is the extended rank likelihood that is defined in equation (2), $P(K|G)$ denotes a prior distribution of a precision matrix K for a given graph structure and $P(G)$ denotes a prior distribution for a graph G.

In particular, we develop a simple and efficient continuous birth–death MCMC algorithm for the posterior computation that converges much faster than the reversible jump MCMC algorithm in Dobra and Lenkoski (2011). Moreover, we evaluate the results by using posterior predictive checks on the scale of the original mixed variables.

For the prior distribution of the graph, we consider a discrete uniform distribution over the graph space, as a non-informative prior. We consider the G-Wishart (Roverato, 2002) distribution as prior distribution of the precision matrix. The G-Wishart distribution is the Wishart distribution restricted to the space of precision matrices with 0 entries specified by a graph G. The G-Wishart density for $K \sim W_G(b, D)$ can be written as

$$P(K|G) = \frac{1}{I_G(b, D)}|K|^{(b-2)/2}\exp\left\{-\frac{1}{2}\mathrm{tr}(DK)\right\},$$

where $b > 2$ are the degrees of freedom, D is a symmetric positive definite matrix and $I_G(b, D)$ is a normalizing constant.

Since the G-Wishart distribution is a conjugate prior together with the multivariate normal distribution, the posterior distribution of K conditionally on graph G is

$$P(K|\mathbf{z}, G) = \frac{1}{I_G(b^*, D^*)}|K|^{(b^*-2)/2}\exp\left\{-\frac{1}{2}\mathrm{tr}(D^*K)\right\},$$

where $b^* = b + n$ and $D^* = D + S$ with $S = \mathbf{z}'\mathbf{z}$ where $\mathbf{z}$ is a single point in $\mathcal{A}(\mathbf{y})$, i.e. a G-Wishart distribution, $W_G(b^*, D^*)$.

2.2.1 Sampling algorithm for posterior inference

Sampling from the joint posterior distribution (3) can be undertaken by a computationally efficient birth–death MCMC algorithm introduced in Mohammadi and Wit (2015) for Gaussian graphical models only. Here we extend their algorithm for the more general case of GCGMs. Our algorithm is based on a continuous time birth–death Markov process in which the algorithm explores the graph space by adding or removing an edge in a birth or death event respectively. The birth and death rates of edges are determined by the stationary distribution of the process. The algorithm is designed in such a way that the stationary distribution equals the target joint posterior distribution of the graph and the precision matrix (3). The time between two successive events has an exponential distribution. Therefore, the probabilities of birth and death events are proportional to their rates.

Mohammadi and Wit (2015), section 3, proved that the birth–death MCMC (BDMCMC) algorithm converges to the target joint posterior distribution of the graph and the precision matrix and proposed to use the birth and death rates

$$\beta_e(K) = \frac{P\{G^{+e}, K^{+e} \setminus (k_{ij}, k_{jj})|Z \in \mathcal{A}(\mathbf{y})\}}{P\{G, K \setminus k_{jj}|Z \in \mathcal{A}(\mathbf{y})\}}, \qquad \text{for each } e \notin E, \tag{4}$$

$$\delta_e(K) = \frac{P\{G^{-e}, K^{-e} \setminus k_{jj}|Z \in \mathcal{A}(\mathbf{y})\}}{P\{G, K \setminus (k_{ij}, k_{jj})|Z \in \mathcal{A}(\mathbf{y})\}}, \qquad \text{for each } e \in E, \tag{5}$$

in which $G^{+e} = (V, E \cup \{e\})$ for the inclusion of an edge from the graph G and K^{+e} represents an updated precision matrix when an edge is included, and similarly for $G^{-e} = (V, E \setminus \{e\})$ and K^{-e}. Details of the BDMCMC algorithm are as follows.

Given a graph $G = (V, E)$ with a precision matrix K, iterate the following steps.

Step 1: sample the latent data. For each $r \in V$ and $j \in \{1, 2, \ldots, n\}$, update the latent value $z_r^{(j)}$ from its full conditional distribution

$$Z_r | K, Z_{V\setminus\{r\}} = z^{(j)}_{K, V\setminus\{r\}} \sim \mathcal{N}\left(-\sum_{r'} \frac{K_{rr'} z^{(j)}_{r'}}{K_{rr}}, \frac{1}{K_{rr}} \right), \tag{6}$$

truncated to the interval in equation (1). This sampling step can be easily modified to handle data that are missing at random, i.e., if y_r is missing, then the full conditional $Z_r | K, Z_{V\setminus\{r\}}$ is the untruncated multivariate normal distribution given in expression (6).

Step 2: sample the graph based on the birth-and-death process.

Step 3: sample the new precision matrix, according to the type of jump.

In our algorithm, the first step is to sample the latent variables given the observed data. Then, on the basis of this sample, we calculate the birth and death rates and waiting times. The birth and death rates are used to calculate the type of jump. Details on how to calculate the birth and death rates efficiently are discussed in Section 2.2.2. Finally in step 3, according to the jump selected, we sample a new precision matrix by using a direct sampling scheme from the G-Wishart distribution that was developed by Lenkoski (2013).

For our algorithm, the Rao–Blackwellized sample mean (Cappé *et al.* (2003), section 2.5) provides an effective way to estimate the posterior probability of each graph. The Rao–Blackwellized estimate of the posterior graph probability is the proportion to the total waiting times for that graph (Fig. 2, bottom right-hand side). The waiting times for each graph act as the weights of that graph (e.g. $\{W_1, W_2, \ldots\}$ in Fig. 2, bottom left-hand side).

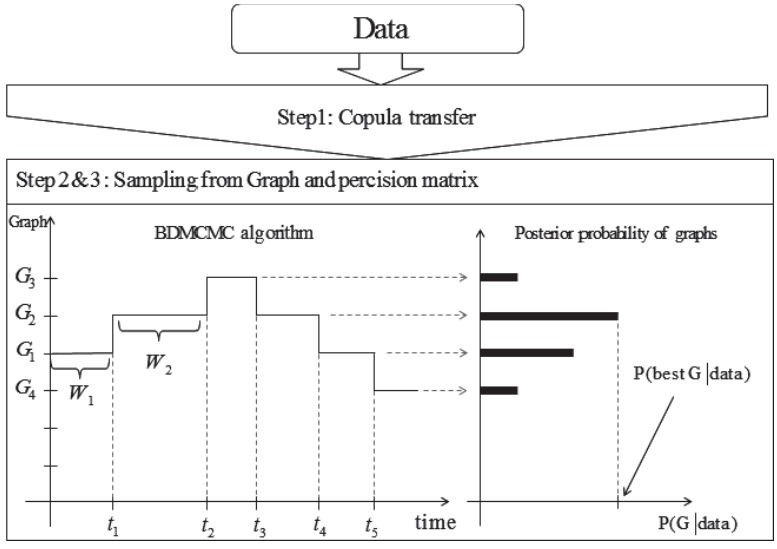


Fig. 2. Visualization of our algorithm: at the top is the mixed observed data transformation using the copula to sample the latent variables; the bottom left-hand side is a continuous time BDMCMC algorithm where $\{W_1, W_2, \ldots\}$ denote waiting times and $\{t_1, t_2, \ldots\}$ denote jumping times; the bottom right-hand side shows the estimated posterior probability of the graphs which are proportional to the sum of their waiting times

2.2.2. Computing the birth and death rates

Calculating the birth and death rates (4) and (5) has been a major bottleneck of the BDMCMC algorithm. Here, we explain how to resolve the computational bottleneck and come up with an efficient way to calculate the death rates; the birth rates are calculated in a similar manner.

Following Mohammadi and Wit (2015) and after some simplification, for each $e = (i, j) \in E$, we have

$$\delta_e(K) = \frac{P(G^{-e})}{P(G)} \frac{I_G(b, D)}{I_{G^{-e}}(b, D)} \left(\frac{D^*_{jj}}{2\pi} \right)^{1/2} H(K, D^*, e), \tag{7}$$

where

$$H(K, D^*, e) = \frac{1}{k_{ii} - k^1_{11}} \exp\left(-\frac{1}{2} \left[\operatorname{tr}\{D^*_{e,e}(K^0 - K^1)\} - \left(D^*_{ii} - \frac{D^{*2}_{ij}}{D^*_{jj}} \right)(k_{ii} - k^1_{11}) \right] \right), \tag{8}$$

in which

$$K^0 = \begin{pmatrix} k_{ii} & 0 \\ 0 & K_{j,V\setminus j}(K_{V\setminus j,\, V\setminus j})^{-1} K_{V\setminus j,\, j} \end{pmatrix},$$

and $K^1 = K_{e,V\setminus e} K^{-1}_{V\setminus e,\, V\setminus e} K_{V\setminus e,e}$. To compute the death rates (7), we need to determine the ratio of the prior normalizing constants, $I_G(b, D)/I_{G^{-e}}(b, D)$.

To compute the ratio of such normalizing constants, Mohammadi and Wit (2015), using ideas from Wang and Li (2012) and Cheng and Lenkoski (2012), developed an approach which borrows ideas from the exchange algorithm (Murray *et al.*, 2012) and the double-Metropolis–Hastings algorithm (Liang, 2010). Following Mohammadi and Wit (2015), we augment the sampling process with an auxiliary variable $\tilde{K} \sim W_G(b, D)$. This results in the death rates

$$\delta_e(K) = \frac{P(G^{-e})}{P(G)} \frac{H(K, D^*, e)}{H(\tilde{K}, D, e)}, \tag{9}$$

in which we evaluated the ratio of the prior normalizing constants with a value of $H(\cdot)$ in equation (8) at $\tilde{K}$.

2.2.3. Simulation study

We perform a simulation study with respect to different graph structures to evaluate the performance of the BDMCMC method proposed and compare it with an alternative approach proposed by Dobra and Lenkoski (2011). We generate data from a latent Gaussian copula model with five different types of variables, including 'Gaussian', 'non-Gaussian', 'ordinal', 'count' and 'binary'. We performed all computations with our R package `BDgraph` (Mohammadi and Wit, 2016a, b).

We consider three different kinds of synthetic graphical model, having p nodes:

(a) *random*—a graph in which the edges are randomly generated from independent Bernoulli distributions with probability $2/(p-1)$;
(b) *cluster*—a graph in which the number of clusters is max $\{2, [p/20]\}$. Each cluster has the same structure as a random graph;
(c) *scale free*—a graph which has a power low degree distribution generated by the Barabási–Albert algorithm (Albert and Barabási, 2002).

With regard to the graph structure G, the corresponding precision matrix is generated from $K \sim W_G(3, \mathbb{I}_p)$. For each graphical model, we consider various scenarios based on two different

Table 1. F_1-score (10) and MSE for our method and the Dobra and Lenkoski (2011) method, with 50 replications and standard deviations in parentheses†

p	n	*Graph*	F_1-*scores*		*MSE*	
			BDMCMC	*Dobra–Lenkoski*	*BDMCMC*	*Dobra–Lenkoski*
10	30	Random	*0.54* (0.13)	0.52 (0.11)	*6.4* (1.4)	9.1 (0.9)
		Cluster	*0.55* (0.16)	0.54 (0.12)	*6.2* (1.5)	8.8 (1.1)
		Scale free	*0.56* (0.17)	0.53 (0.10)	*5.9* (1.4)	8.9 (0.8)
10	100	Random	*0.69* (0.14)	0.67 (0.12)	*4.5* (1.7)	6.3 (1.4)
		Cluster	*0.73* (0.14)	0.72 (0.12)	*3.9* (1.5)	5.5 (1.1)
		Scale free	*0.68* (0.15)	0.67 (0.13)	*4.3* (1.5)	6.1 (1.0)
40	400	Random	*0.74* (0.07)	0.55 (0.05)	*19.3* (4.8)	60.5 (6.0)
		Cluster	*0.80* (0.05)	0.65 (0.05)	*19.6* (4.7)	59.3 (7.0)
		Scale free	*0.73* (0.08)	0.52 (0.07)	*19.5* (5.4)	64.8 (12.4)
40	800	Random	*0.83* (0.06)	0.70 (0.07)	*11.8* (3.5)	35.9 (6.5)
		Cluster	*0.84* (0.05)	0.75 (0.05)	*16.7* (4.7)	37.2 (5.8)
		Scale free	*0.82* (0.07)	0.68 (0.07)	*12.5* (4.1)	37.6 (6.7)

†The best models are in italics.

numbers of variables $p = \{10, 40\}$ and various sample sizes; Table 1. For each scenario, we generate data and fit our BDMCMC and the Dobra and Lenkoski (2011) approaches using a uniform prior for the graph and the G-Wishart prior $W_G(3, \mathbb{I}_p)$ for the precision matrix. We run the two algorithms with the same starting points using 100000 iterations and 50000 iterations as a burn-in.

To assess the performance of the graph structure, we compute the F_1-score measure which is defined as

$$F_1\text{-score} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}, \tag{10}$$

where TP, FP and FN are the number of true positive, false positive and false negative results respectively. The F_1-score lies between 0 and 1, where 1 denotes perfect identification and 0 denotes poor identification. Also, we use the mean-square error (MSE) of the posterior edge inclusion probabilities. We calculate the posterior edge inclusion probabilities on the basis of Rao–Blackwellization (Cappé *et al.* (2003), section 2.5) for each possible edge $e = (i, j)$ as

$$\hat{p}_e = \frac{\sum_{t=1}^{N} I(e \in G^{(t)})\, W(K^{(t)})}{\sum_{t=1}^{N} W(K^{(t)})}, \tag{11}$$

where N is the number of iterations and $W(K^{(t)})$ is the waiting time for the graph $G^{(t)}$ with the precision matrix $K^{(t)}$.

Table 1 reports comparisons of the BDMCMC method with the Dobra and Lenkoski (2011) method where we repeat the experiments 50 times and report the average F_1-score and MSE with their standard errors in parentheses. Our BDMCMC method performs well overall as its F_1-score is larger and its MSE is lower than those of the Dobra and Lenkoski (2011) method in all scenarios under consideration, mainly because of its faster convergence rate. Simulations suggest that the BDMCMC algorithm converges in approximately a quarter of the time of the

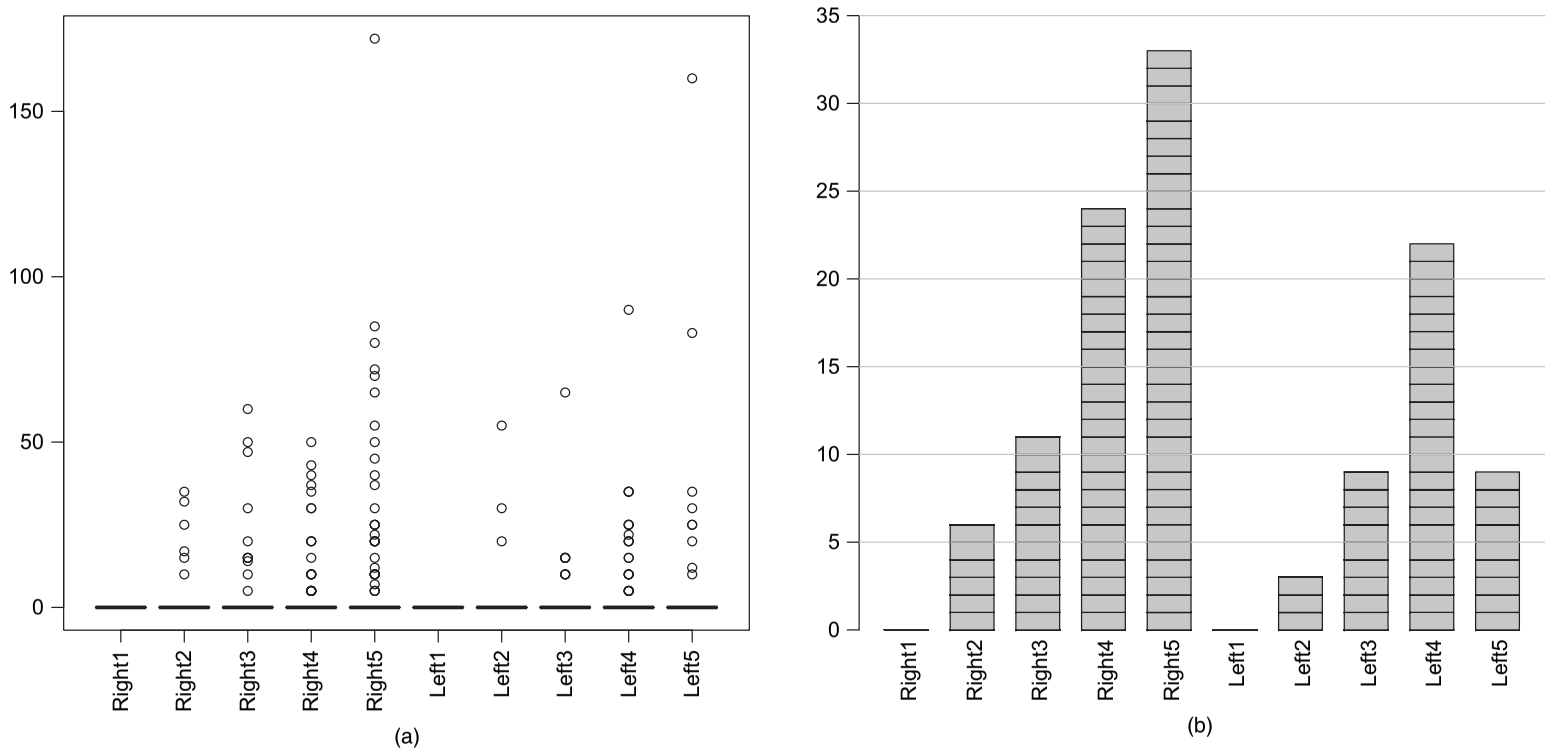


Fig. 3. (a) Boxplot of angles of the 10 fingers of all 279 patients and (b) frequency histogram of rays affected with Dupuytren disease for all 10 fingers

algorithm of Dobra and Lenkoski, which results in the different F_1-score and MSE behaviour under finite MCMC iterations, as reported in Table 1.

3. Analysis of Dupuytren disease data

Here we analyse the data that were collected on patients who have Dupuytren disease in both hands from the north of the Netherlands by the Department of Plastic Surgery of the University Medical Centre Groningen. The data were originally described by Lanting *et al.* (2013) and Lanting, Nooraee, Werker and van den Heuvel (2014). The data consist of 279 patients who have Dupuytren disease ($n = 279$); among those patients, 79 have an irreversible flection contracture in at least one of their fingers. Therefore, the data consist of many 0s as shown in Fig. 3, i.e., though the hands are affected by the disease, the fingers did not show any sign of contraction and the total angle measure is taken as 0.

The severity of the disease in all 10 fingers of the patients is measured by the angle of each finger, which is defined as the sum of angles for the metacarpophalangeal joints. To study the potential phenotype risk factors of Dupuytren disease, we consider 13 potential phenotype risk factors. These are smoking habits (Smoking), alcohol consumption (Alcohol), whether

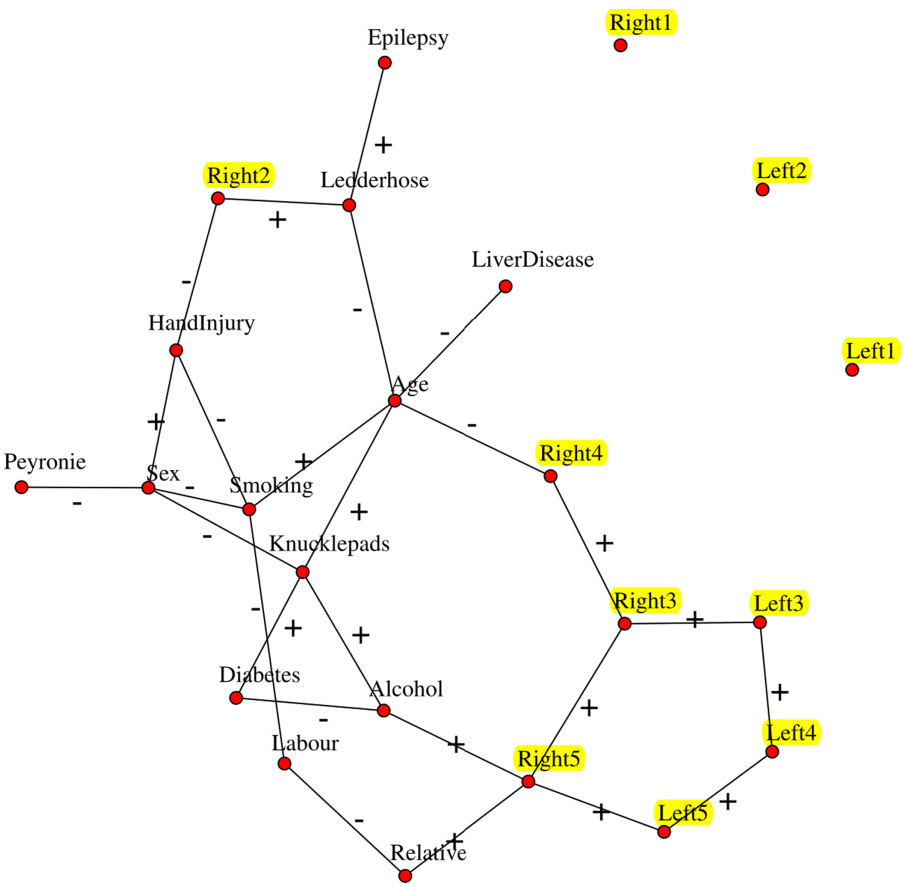


Fig. 4. Inferred graph for the Dupuytren disease data set based on 13 risk factors and the total degrees of flection in all 10 fingers: it reports the selected graph with 26 edges for which their posterior inclusion probabilities (11) are more than 0.4 (+, positive relationship between nodes; −, negative relationship)

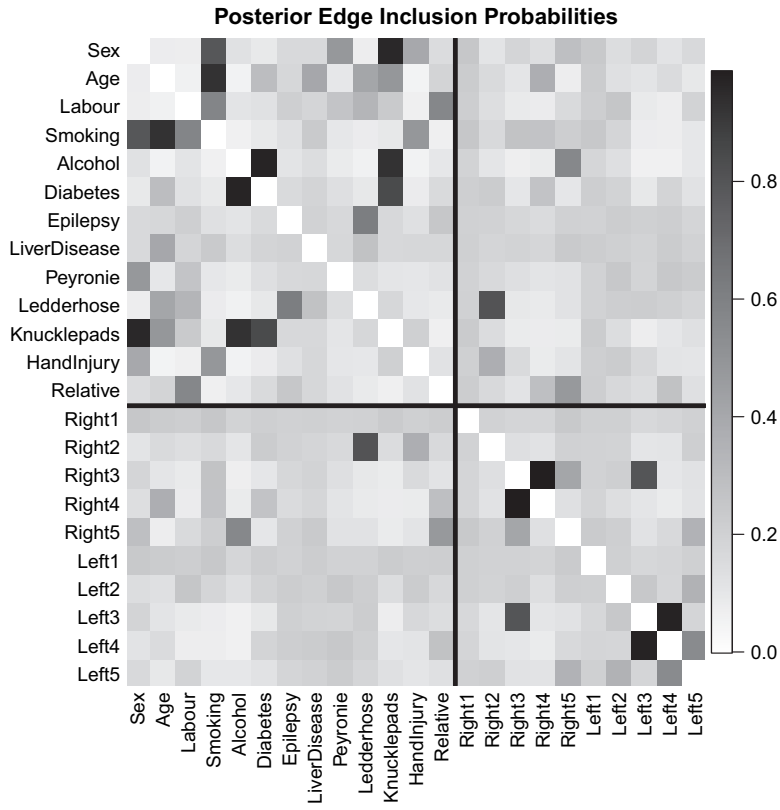


Fig. 5. Image visualization of the posterior edge inclusion probabilities of all possible edges in the graph, for 10 fingers with 13 risk factors

participants performed manual labour during a significant part of their life (Labour), whether they had sustained hand injury in the past including surgery (HandInjury), disease history information about the presence of factors Ledderhose, Diabetes, Epilepsy, Peyronie, Knucklepads and liver disease (LiverDisease) and familial occurrence of Dupuytren disease which is defined as a first-degree relative with Dupuytren disease (Relative).

For each finger we measure angles of the metacarpophalangeal joints, three interphalangeal joints (for thumbs we measure only two interphalangeal joints); then we sum those angles for each finger as a measure of the severity of Dupuytren disease. The total angles can vary from 0° to 270°; however, in this data set the minimum is 0° and the maximum is 157°. The age of participants (in years) ranges from 40 to 89 years, with an average age of 66 years. Smoking is binned into three ordered categories (never, stopped and smoking). Amount of alcohol consumption is binned into eight ordered categories (ranging from no alcohol to more than 20 units of consumption per week). All other variables are binary.

In Section 3.1, we infer the Dupuytren disease network of the fingers with the 13 potential risk factors on the basis of the BDMCMC approach. In Section 3.2, we consider only the severity measurements of the 10 fingers to infer the interaction between the fingers.

3.1. Inference for Dupuytren disease with risk factors

We apply our Bayesian framework to infer the conditional (in)dependence structure among the

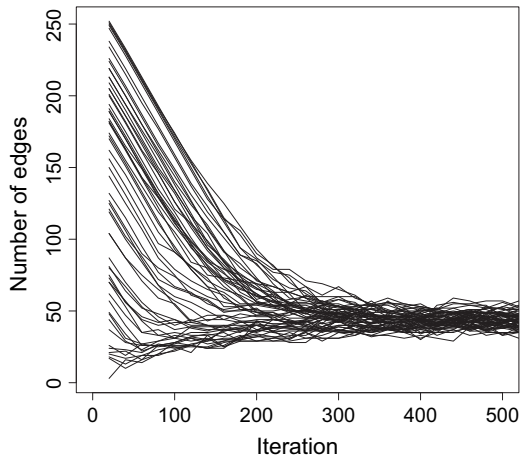


Fig. 6. Trace plot of the number of edges included in the estimated graphs against iterations of the BDMCMC algorithm with 100 different starting points

23 variables, to identify the potential risk factors of Dupuytren disease and to discover how they affect the disease. In implementing the BDMCMC approach to analyse this data set, we place a G-Wishart prior $W_G(3, \mathbb{I}_{23})$ on the precision matrix. We run the BDMCMC algorithm for 2 million iterations with 1 million sweeps burn-in. The results are displayed in Figs 4 and 5.

By using the median probability model of Barbieri and Berger (2004), Fig. 4 visualizes 26 edges that have posterior probabilities larger than 0.4. Similarly, Fig. 5 shows the image of all posterior inclusion probabilities where the degree of darkness increases with increasing posterior probabilities.

The edges in the graph show the interactions between the 10 severity measurements of Dupuytren disease and 13 risk factors. For example, the results show that factors Age, Alcohol, Ledderhose, HandInjury and Relative, among those 13 risk factors, have a significant association with the severity of Dupuytren disease. As expected, increased alcohol use, the presence of ledderhose disease and having a direct relative with Dupuytren disease increase the severity of the disease. However, surprisingly, correcting for all other variables, age and the presence of hand injury decreases the severity of Dupuytren disease. Fig. 4 also shows that factor Age is a hub in this graph and it plays a significant role as it affects the severity of the disease directly and indirectly through the influence of other risk factors such as Ledderhose.

Further we checked the stability of the graph selected with highest posterior probability at convergence of the algorithm with 100 different starting points. The resulting Fig. 6 shows the traces of the number of edges in the estimated graphs plotted against iterations of the BDMCMC algorithm with the 100 different starting points. The plot shows good mixing around a stable graph model size, which is 42, and the algorithm converges after around 300 iterations.

3.2. Severity of Dupuytren disease between pairs of fingers

In this section, we consider the relationship between the occurrence of Dupuytren disease in pairs of fingers on both hands. Interaction between fingers is important because it help surgeons to decide whether they should operate on one finger or on multiple fingers simultaneously. The main idea is that, if fingers are almost independent in terms of the severity of Dupuytren disease, there is no reason to operate on the fingers simultaneously. In contrast, if there is a strong

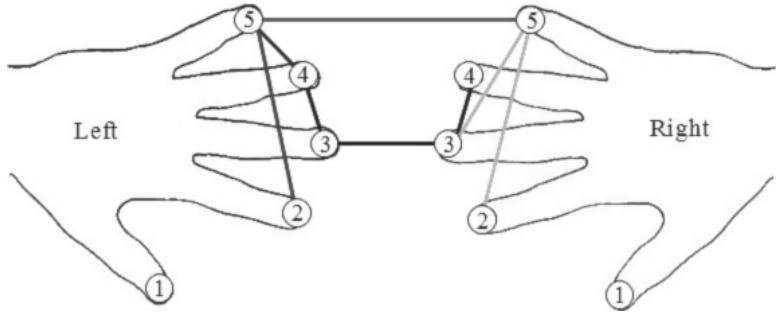


Fig. 7. Inferred graph of the Dupuytren disease data set based on the total degrees of flection in all 10 fingers: it reports the graph selected with eight edges for which their posterior inclusion probabilities (11) are more than 0.4

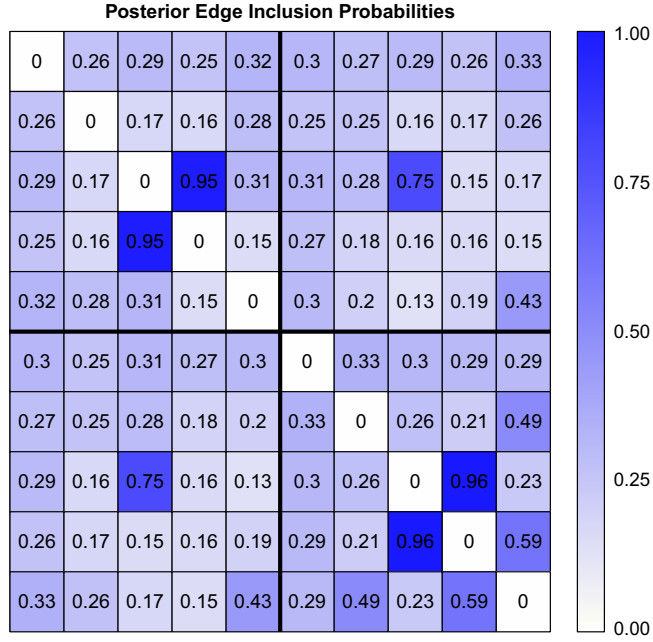


Fig. 8. Image visualization of the posterior edge inclusion probabilities of all possible edges in the graph, for 10 fingers

relationship between fingers, then joint surgery may be recommended if one of the fingers is affected.

We apply the BDMCMC approach for the 10 variables of Dupuytren disease severity measures by using the G-Wishart $W_G(3, \mathbb{I}_{10})$ prior on the precision matrix. We run the BDMCMC algorithm for 2 million iterations with 1 million sweeps as burn-in. The results are displayed in Fig. 7 and Fig. 8.

Fig. 7 visualizes the graph selected with eight edges, for which the posterior inclusion probabilities (11) are greater than 0.4. The edges in the graph show the interactions between the fingers with regard to the severity of Dupuytren disease.

The results show significant co-occurrences of Dupuytren disease in the ring fingers and middle fingers in both hands. This suggests that disease in the middle finger is strongly associated

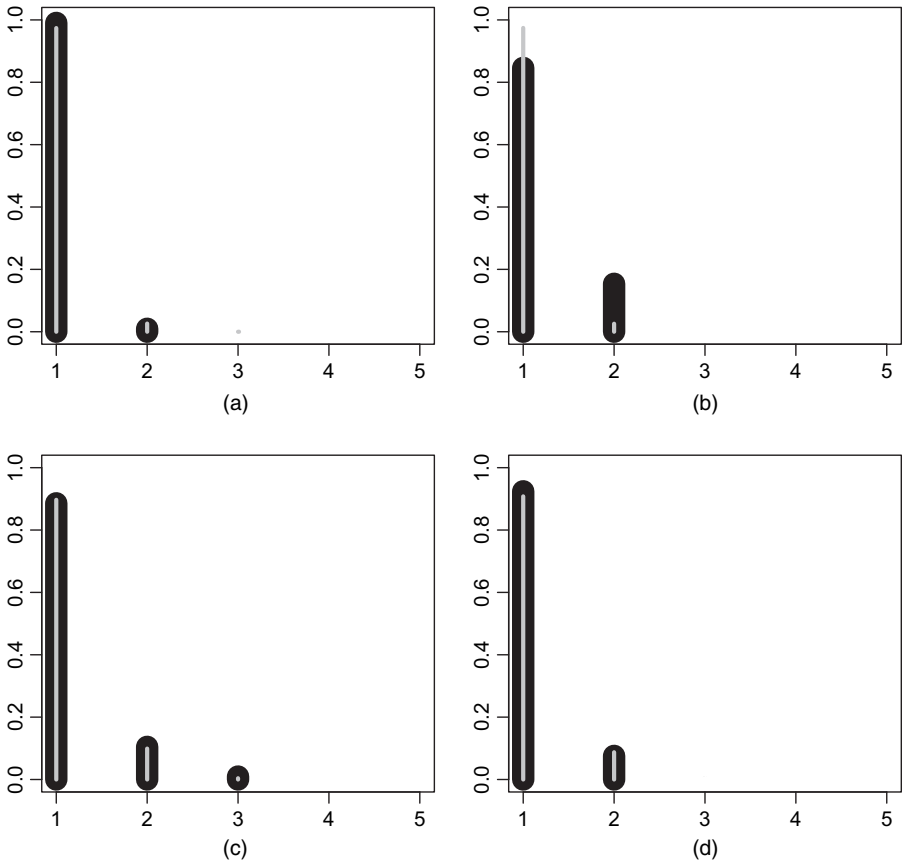


Fig. 9. Empirical (▮) and predictive (▮) conditional distributions for the total angle of finger 4 in the right hand conditionally on four categories of variable Age: (a) $P(\text{Right 4} \mid 40 < \text{age} < 50)$; (b) $P(\text{Right 4} \mid 50 < \text{age} < 60)$; (c) $P(\text{Right 4} \mid 60 < \text{age} < 70)$; (d) $P(\text{Right 4} \mid 70 < \text{age} < 90)$

with the ulnar side of the hand. Surprisingly, our results also show a strong relationship between the middle fingers in both hands. Moreover, the results show that the joint interactions between fingers in both hands are almost symmetric. These results support the hypotheses that the disease has genetic factors or other biological factors that affect similar fingers in both hands.

3.3. Fit of model to Dupuytren data

Posterior predictive checks can be used for checking whether the Bayesian approach fits the Dupuytren data set well or not. If the model fits the Dupuytren data, then simulated data that are generated under the model should look like the observed data. In this regard, first, on the basis of our estimated graph from the BDMCMC algorithm in Section 3.1, we draw simulated data from the posterior predictive distribution. Then, we compare the samples with our observed data. Any systematic differences between the simulations and the data determine potential failings of the model.

We obtain the conditional distributions of the potential risk factors and disease severity measures on the fingers for both simulated and observed data. The empirical and predictive conditional distributions of some selected variables are presented in Figs 9, 10 and 11.

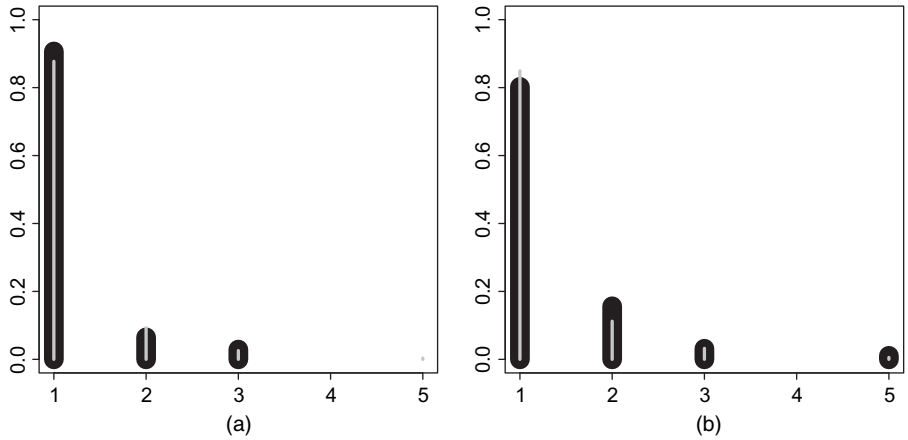


Fig. 10. Empirical (▮) and predictive (▯) conditional distributions for total angles of finger 5 in the right hand conditionally on variable Relative: (a) $P(\text{Right 5}|\text{Relative} = 0)$; (b) $P(\text{Right 5}|\text{Relative} = 1)$

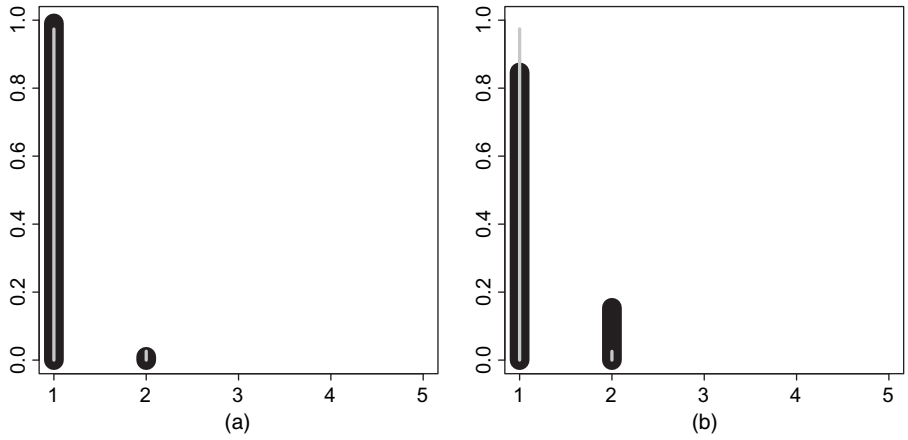


Fig. 11. Empirical (▮) and (▯) predictive conditional distributions for total angles of finger 2 in the right hand conditionally on variable Ledderhose: (a) $P(\text{Right 2}|\text{Ledderhose} = 0)$; (b) $P(\text{Right 2}|\text{Ledderhose} = 1)$

Fig. 9 displays the empirical and predictive distributions of disease severity measure on finger 4 in the right hand (Right4) conditionally on variable Age in four categories $\{(40, 50), (50, 60), (60, 70), (70, 90)\}$. The variable Right4, based on the Tubiana classification, was grouped into five categories: category 1, 0° for the total angle; 2, degree between $(1°, 45°)$; 3, degree between $(46°, 90°)$; 4, degree between $(90°, 135°)$; 5, degree more than 135°. The results in Fig. 9 show that the fit is good, since the predicted conditional distributions, in general, are the same as the empirical distributions.

Similarly, Fig. 10 plots the empirical and predictive distribution of disease severity measure on finger 5 in the right hand (Right5) conditionally on variable Relative and Fig. 11 plots the empirical and predictive distribution of disease severity measure on finger 2 in the right hand (Right2) conditionally on variable Ledderhose. These results also suggest that the BDMCMC approach fits the Dupuytren data well as the predicted conditional distributions are in agreement with the empirical distributions.

4. Conclusion

In this paper we have implemented a Bayesian method for discovering the effect of potential risk factors of Dupuytren disease and the underling relationships between fingers on both hands with regard to severity of the disease.

The results of the case-study clearly demonstrate that age, alcohol, relative and ledderhose diseases all affect the severity of Dupuytren disease directly. However, in contrast with what was reported before, age and the presence of hand injury are inversely related to the severity of Dupuytren disease when correcting for the other variables. Other risk factors affect Dupuytren disease only indirectly. Another important result is that the severity of Dupuytren disease in fingers is correlated: in particular, the middle finger with the ring finger. This implies that a surgical intervention on either the ring or the middle finger should preferably be executed simultaneously.

In our case-study, we consider the 13 potential phenotype risk factors. It would be interesting to consider those phenotype risk factors jointly with other genotype risk factors that are cited in the literature. For example, in a genomewide association study, nine genes were identified as being associated with Dupuytren disease (Dolmans *et al*., 2011). Bayesian inference for all these risk factors requires a computationally efficient search algorithm that can potentially explore the underlying graph structure to uncover complicated patterns among these variables. Our Bayesian framework is well suited to this kind of problem.

We have compared our BDMCMC Bayesian approach with an alternative Bayesian approach (Dobra and Lenkoski, 2011) by using a simulation study on various types of graph structures. Although both approaches converge to the same posterior distribution our approach has some clear advantages on finite MCMC runs. This difference is mainly due to our implementation of a computationally efficient algorithm, which is a continuous time MCMC algorithm based on a birth–death process.

Of course, our extended Bayesian method is not limited only to this type of data. It can potentially be applied to any kind of data where the observed variables are binary, ordinal or continuous.

Acknowledgements

The authors thank Paul Werker and Dieuwke Broeksma (University Medical Centre Groningen) for providing the pictures of the affected hand, as well as the patient for agreeing to use these pictures for this publication.

References

Abegaz, F. and Wit, E. (2015) Copula Gaussian graphical models with penalized ascent Monte Carlo EM algorithm. *Statist. Neerland.*, **69**, 419–441.

Albert, R. and Barabási, A.-L. (2002) Statistical mechanics of complex networks. *Rev. Mod. Phys.*, **74**, 47–97.

Barbieri, M. M. and Berger, J. O. (2004) Optimal predictive model selection. *Ann. Statist.*, **32**, 870–897.

Bayat, A. and McGrouther, D. (2006) Management of Dupuytren's disease—clear advice for an elusive condition. *Ann. R. Coll. Surg. Engl.*, **88**, 3–8.

Cappé, O., Robert, C. P. and Rydén, T. (2003) Reversible jump, birth-and-death and more general continuous time Markov chain Monte Carlo samplers. *J. R. Statist. Soc.* B, **65**, 679–700.

Cheng, Y. and Lenkoski, A. (2012) Hierarchical Gaussian graphical models: beyond reversible jump. *Electron. J. Statist.*, **6**, 2309–2331.

Dobra, A. and Lenkoski, A. (2011) Copula Gaussian graphical models and their application to modeling functional disability data. *Ann. Appl. Statist.*, **5**, 969–993.

Dobra, A., Lenkoski, A. and Rodriguez, A. (2011) Bayesian inference for general Gaussian graphical models with application to multivariate lattice data. *J. Am. Statist. Ass.*, **106**, 1418–1433.

Dolmans, G. H., Werker, P. M., Hennies, H. C., Furniss, D., Festen, E. A., Franke, L., Becker, K., van der Vlies, P., Wolffenbuttel, B. H., Tinschert, S., Toliat, M. R., Nothnagel, M., Franke, A., Klopp, N., Wichmann, H.-E., Nürnberg, P., Giele, H., Ophoff, R. A. and Wijmenga, C. (2011) Wnt signaling and Dupuytren's disease. *New Engl. J. Med.*, **365**, 307–317.

Geoghegan, J., Forbes, J., Clark, D., Smith, C. and Hubbard, R. (2004) Dupuytren's disease risk factors. *J. Hnd Surg.*, **29**, 423–426.

Godtfredsen, N. S., Lucht, H., Prescott, E., Sørensen, T. I. and Grønbæk, M. (2004) A prospective study linked both alcohol and tobacco to Dupuytren's disease. *J. Clin. Epidem.*, **57**, 858–863.

Hoff, P. D. (2007) Extending the rank likelihood for semiparametric copula estimation. *Ann. Appl. Statist.*, **1**, 265–283.

Lanting, R., Broekstra, D. C., Werker, P. M. and van den Heuvel, E. R. (2014) A systematic review and meta-analysis on the prevalence of Dupuytren disease in the general population of western countries. *Plast. Reconstr. Surg.*, **133**, 593–603.

Lanting, R., den Heuvel, E. R., Westerink, B. and Werker, P. M. (2013) Prevalence of Dupuytren disease in the Netherlands. *Plast. Reconstr. Surg.*, **132**, 394–403.

Lanting, R., Nooraee, N., Werker, P. and van den Heuvel, E. (2014) Patterns of Dupuytren disease in fingers; studying correlations with a multivariate ordinal logit model. *Plast. Reconstr. Surg.*, **134**, 483–490.

Lauritzen, S. (1996) *Graphical Models*. New York: Oxford University Press.

Lenkoski, A. (2013) A direct sampler for G-Wishart variates. *Stat*, **2**, 119–128.

Liang, F. (2010) A double Metropolis-Hastings sampler for spatial models with intractable normalizing constants. *J. Statist. Computn Simuln*, **80**, 1007–1022.

Liu, H., Han, F., Yuan, M., Lafferty, J. and Wasserman, L. (2012) High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.*, **40**, 2293–2326.

Meinshausen, N. and Bühlmann, P. (2006) High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, **34**, 1436–1462.

Meyerding, H. W. (1936) Dupuytren's contracture. *Arch. Surg.*, **32**, 320–333.

Mohammadi, A. and Wit, E. C. (2015) Bayesian structure learning in sparse Gaussian graphical models. *Baysn Anal.*, **10**, 109–138.

Mohammadi, A. and Wit, E. C. (2016a) BDgraph: an R package for Bayesian structure learning in graphical models. *J. Statist. Softwr.*, to be published.

Mohammadi, A. and Wit, E. (2016b) BDgraph: Bayesian graph selection based on birth-death MCMC approach. *R Package Version 2.27*. University of Groningen, Groningen.

Murray, I., Ghahramani, Z. and MacKay, D. (2012) MCMC for doubly-intractable distributions. *Preprint arXiv:1206.6848*.

Roverato, A. (2002) Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scand. J. Statist.*, **29**, 391–411.

Shih, B. and Bayat, A. (2010) Scientific understanding and clinical management of Dupuytren disease. *Nat. Rev. Rheum.*, **6**, 715–726.

Tubiana, R., Simmons, B. and DeFrenne, H. (1982) Location of Dupuytren's disease on the radial aspect of the hand. *Clin. Orthpaed. Reltd Res.*, no. 168, 222–229.

Wang, H. and Li, S. (2012) Efficient Gaussian graphical model determination under G-Wishart prior distributions. *Electron. J. Statist.*, **6**, 168–198.

Whittaker, J. (2009) *Graphical Models in Applied Multivariate Statistics*. Chichester: Wiley.

Yuan, M. and Lin, Y. (2007) Model selection and estimation in the Gaussian graphical model. *Biometrika*, **94**, 19–35.